Order statistics for histogram data and a box plot visualization tool

Rosanna Verde*, Antonio Balzanella*, Antonio Irpino*

*Second University of Naples, Caserta, Italy
rosanna.verde@unina2.it, antonio.balzanella@unina2.it, antonio.irpino@unina2.it

Abstract. This paper deals with new descriptive statistics for histogram data, in the framework of symbolic data analysis. A main contribution consists in defining the main order statistics (median and quartiles) of a histogram variable using the quantile functions associated with the corresponding empirical distribution functions of the observed histograms. The definition of an order relationship between quantile functions is based on an appropriate probabilistic metric: the ℓ^p Wasserstein distance. Starting from the median and quartile functions definition, we extend the classic box-plot representation for set of quantile functions. Finally, we propose new measures of variability and skewness for a histogram variable associated with this representation. An application on real data allows us to corroborate the proposed measures and the new box-plot visualization tool.

1 Introduction

The advance of technology is making possible to observe and to collect very large datasets. The analysis of such data is often performed after a summarization step whose aims are to obtain a more manageable information, in size and in terms of computational resources, while preserving as much as possible the information of the entire data set. The representation of data through histograms is a common practice in data summarization. In fact, a histogram is parsimonious representation, with respect to storage requirements, and it provides an idea of the underlying distribution of the observed data or of subsets of values observed for a single attribute.

Symbolic Data Analysis (in short SDA) (Boch and Diday, 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture, 2008) provides a formalization of a new symbolic descriptor, the *histogram variable* which is a particular case of symbolic multi-valued modal variable. Several techniques (Clustering, Regression, PCA,...) have been proposed in Billard and Diday (2003) to analyze histogram data. Some basic statistics like the *sample mean* and the *standard deviation* for a histogram variable have been introduced in Bertrand and Goupil (2000), Billard and Diday (2003), Billard and Diday (2006) and Irpino et al. (2006). Graphical tools for visualizing symbolic data (including histogram data) have been also presented by Noirhomme-Fraiture and Rouard (1997) and by Noirhomme-Fraiture and Nahimana (2008).